Discussion of The Role of SkewSymmetric Distributions in Bayesian Inference: Conjugacy, Scalable Approximations and Asymptotics

Mengyang Gu

Department of Statistics and Applied Probability
University of California, Santa Barbara

Background: Bayesian probit regression

- (**Model**). Given independent binary data $y_1, ..., y_n$ from a probit regression model $y_i \mid \boldsymbol{\beta} \sim Bern[\Phi(\mathbf{x}_i^T\boldsymbol{\beta})]$, for $i = 1, ..., n$ with prior $\boldsymbol{\beta} \sim \mathrm{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$ and Φ denoting the cumulative distribution function (CDF) of a standard normal distribution.
- (**Posterior**.) Denoting ϕ_p the density of zero mean normal distribution with variance $\boldsymbol{\Omega}$, we have

$$p(\boldsymbol{\beta} \mid \mathbf{y}) = \frac{\phi_p(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega}) \prod_{i=1}^n \Phi(\mathbf{x}_i^T\boldsymbol{\beta})^{y_i}(1 - \Phi(\mathbf{x}_i^T\boldsymbol{\beta}))^{1-y_i}}{\int_{\mathbb{R}^p} \phi_p(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega}) \prod_{i=1}^n \Phi(\mathbf{x}_i^T\boldsymbol{\beta})^{y_i}(1 - \Phi(\mathbf{x}_i^T\boldsymbol{\beta}))^{1-y_i} d\boldsymbol{\beta}}$$

- (**Question**.) Markov Chain Monte Carlo (MCMC) sampling is slow.
  Q1: Do we have a conjugate prior? **Q2**: Is the computation scalable?
  Q3: Can we extend the results to other relevant models?

Conjugacy by the unified skewed-normal distribution (SUN)

- Denoting the SUN density $\boldsymbol{\beta} \sim \text{SUN}_{p,q}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ with $\boldsymbol{\xi} \in \mathbb{R}^p$, $\boldsymbol{\Omega} \in \mathbb{R}^q$, $\boldsymbol{\Delta} \sim \mathbb{R}^{p,q}$, $\boldsymbol{\gamma} \in \mathbb{R}^q$, and $\boldsymbol{\Gamma} \in \mathbb{R}^{n,n}$ full rank matrix [Chen et al., 2016], with density

$$p(\boldsymbol{\beta} \mid \boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma}) = \phi_p(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega}) \frac{\Phi_q(\boldsymbol{\gamma} + \boldsymbol{\Delta}^T \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1} (\boldsymbol{\beta} - \boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^T \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta})}{\Phi_q(\boldsymbol{\gamma}; \boldsymbol{\Gamma})} \quad (1)$$

  where $\boldsymbol{\Omega} = \boldsymbol{\omega} \bar{\boldsymbol{\Omega}} \boldsymbol{\omega}$ is a covariance matrix, with $\bar{\boldsymbol{\Omega}}$ being a correlation matrix and $\boldsymbol{\omega}$ being a diagonal matrix for squared root of the diagonal values of $\boldsymbol{\Omega}$.

- For the probit model, $y_i \mid \boldsymbol{\beta} \sim Bern[\Phi(\mathbf{x}_i^T \boldsymbol{\beta})]$, for $i = 1, ..., n$ with prior $\boldsymbol{\beta} \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$, the posterior follows [Durante, 2019]

$$\boldsymbol{\beta} \mid \mathbf{y} \sim \text{SUN}_{p,q}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \bar{\boldsymbol{\Omega}} \boldsymbol{\omega} \mathbf{D}^T \mathbf{s}^{-1}, \mathbf{s}^{-1} \mathbf{D} \boldsymbol{\xi}, \mathbf{s}^{-1} (\mathbf{D} \boldsymbol{\Omega} \mathbf{D}^T + \mathbf{I}_n) \mathbf{s}^{-1}),$$

  where a $n \times p$ matrix $\mathbf{D} = \text{diag}(2y_1 - 1, ..., 2y_n - 1)\mathbf{X}$ and a $n \times n$ diagonal matrix $\mathbf{s} = \left[(\mathbf{D} \boldsymbol{\Omega} \mathbf{D}^T + \mathbf{I}_n) \odot \mathbf{I}_n\right]^{1/2}$ with $\odot$ denoting the elementwise product.

Conjugacy by the unified skewed-normal distribution (SUN)

- Denoting the SUN density $\boldsymbol{\beta} \sim \text{SUN}_{p,q}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma})$ with $\boldsymbol{\xi} \in \mathbb{R}^p$, $\boldsymbol{\Omega} \in \mathbb{R}^q$, $\boldsymbol{\Delta} \sim \mathbb{R}^{p,q}$, $\boldsymbol{\gamma} \in \mathbb{R}^q$, and $\boldsymbol{\Gamma} \in \mathbb{R}^{n,n}$ full rank matrix [Chen et al., 2016], with density

$$p(\boldsymbol{\beta} \mid \boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \boldsymbol{\Gamma}) = \phi_p(\boldsymbol{\beta} - \boldsymbol{\xi}; \boldsymbol{\Omega}) \frac{\Phi_q(\boldsymbol{\gamma} + \boldsymbol{\Delta}^T \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\omega}^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi}); \boldsymbol{\Gamma} - \boldsymbol{\Delta}^T \bar{\boldsymbol{\Omega}}^{-1} \boldsymbol{\Delta})}{\Phi_q(\boldsymbol{\gamma}; \boldsymbol{\Gamma})} \quad (1)$$

  where $\boldsymbol{\Omega} = \boldsymbol{\omega} \bar{\boldsymbol{\Omega}} \boldsymbol{\omega}$ is a covariance matrix, with $\bar{\boldsymbol{\Omega}}$ being a correlation matrix and $\boldsymbol{\omega}$ being a diagonal matrix for squared root of the diagonal values of $\boldsymbol{\Omega}$.

- For the probit model, $y_i \mid \boldsymbol{\beta} \sim Bern[\Phi(\mathbf{x}_i^T \boldsymbol{\beta})]$, for $i = 1, ..., n$ with prior $\boldsymbol{\beta} \sim \text{N}_p(\boldsymbol{\xi}, \boldsymbol{\Omega})$, the posterior follows [Durante, 2019]

$$\boldsymbol{\beta} \mid \mathbf{y} \sim \text{SUN}_{p,q}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \bar{\boldsymbol{\Omega}} \boldsymbol{\omega} \mathbf{D}^T \mathbf{s}^{-1}, \mathbf{s}^{-1} \mathbf{D} \boldsymbol{\xi}, \mathbf{s}^{-1}(\mathbf{D} \boldsymbol{\Omega} \mathbf{D}^T + \mathbf{I}_n) \mathbf{s}^{-1}),$$

  where a $n \times p$ matrix $\mathbf{D} = \text{diag}(2y_1 - 1, ..., 2y_n - 1)\mathbf{X}$ and a $n \times n$ diagonal matrix $\mathbf{s} = \left[(\mathbf{D} \boldsymbol{\Omega} \mathbf{D}^T + \mathbf{I}_n) \odot \mathbf{I}_n\right]^{1/2}$ with $\odot$ denoting the elementwise product.

- **Answer to Q1**: Yes, we have conjugacy.
- Many nice properties: e.g. normalizing constant and mode of posteriors of SUN can be computed; sampling distribution can be constructed; predictive distributions, linear combination, and conditional distributions are all SUN.

Computational challenge and data augmentation

- **Computational challenge**. The SUN density involves computing a CDF of multivariate normal of n dimensions which may contain $\mathcal{O}(n^3)$ operations (due to computing the Cholesky factor of the covariance). Furthermore, sampling requires n-variate truncated normals [Botev, 2017].

Computational challenge and data augmentation

- **Computational challenge**. The SUN density involves computing a CDF of multivariate normal of n dimensions which may contain $\mathcal{O}(n^3)$ operations (due to computing the Cholesky factor of the covariance). Furthermore, sampling requires n-variate truncated normals [Botev, 2017].
- **Data augmentation**. For probit regression models, data augmentation [Albert and Chib, 1993] has been widely used in MCMC and variational Bayes:

$$y_i = \mathbb{1}_{z_i > 0},\ (z_i \mid \boldsymbol{\beta}) \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, 1),\ \text{and}\ \boldsymbol{\beta} \sim N_p(\mathbf{0}, \nu_p^2 \mathbf{I}_p),$$

where the conditional posterior follows

$$(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y}) = N_p(\mathbf{V}\mathbf{X}^T\mathbf{z}, \mathbf{V}), \mathbf{V} = (\nu_p^{-2}\mathbf{I}_p + \mathbf{X}^T\mathbf{X})^{-1},$$

$$(z_i \mid \boldsymbol{\beta}, \mathbf{z}_{-i}, \mathbf{y}) \sim \begin{cases} \text{TN}[\mathbf{x}_i^T\boldsymbol{\beta}, 1, (0, +\infty)], & \text{if } y_i = 1, \\ \text{TN}[\mathbf{x}_i^T\boldsymbol{\beta}, 1, (-\infty, 0)], & \text{if } y_i = 0, \end{cases}$$

for $i = 1, ..., n$. **Note**: it may need $\mathcal{O}(p^3)$ operations for factorization/inversion of a $p \times p$ matrix (supposing $n > p$ and $\mathbf{V}$ is full rank), but it only needs to be done once.

Variational Bayes (VB)

One answer to Q2:

- **(VB with mean field family)**. Consider mean field family $q_{MF}(\mathbf{z}, \boldsymbol{\beta}) = q(\mathbf{z})q(\boldsymbol{\beta})$. Then maximize

$$\text{ELBO}[q_{MF}(\boldsymbol{\beta}, \mathbf{z})] := -KL[q_{MF}(\boldsymbol{\beta}, \mathbf{z})||p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})] + c,$$

through iterations [Blei et al., 2017]

$$q_{MF}^{(t)}(\boldsymbol{\beta}) = \exp\left[E_{q_{MF}^{(t-1)}(z)}\{\log p(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y})\}\right], q_{MF}^{(t)}(\boldsymbol{\beta}) = \exp\left[E_{q_{MF}^{(t-1)}(z)}\{\log p(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y})\}\right]$$

which approximate $p(\boldsymbol{\beta}, \mathbf{z}, \mid y)$ via a multivariate Gaussian $q_{MF}^{*}(\boldsymbol{\beta})$ and a product of truncated normals $\prod_{i=1}^{n} q_{MF}^{*}(z_i)$.

Variational Bayes (VB)

One answer to Q2:

- **(VB with mean field family)**. Consider mean field family $q_{MF}(\mathbf{z}, \boldsymbol{\beta}) = q(\mathbf{z})q(\boldsymbol{\beta})$. Then maximize

$$\text{ELBO}[q_{MF}(\boldsymbol{\beta}, \mathbf{z})] := -KL[q_{MF}(\boldsymbol{\beta}, \mathbf{z})||p(\boldsymbol{\beta}, \mathbf{z} \mid \mathbf{y})] + c,$$

  through iterations [Blei et al., 2017]

$$q^{(t)}_{MF}(\boldsymbol{\beta}) = \exp\left[E_{q^{(t-1)}_{MF}(z)}\{\log p(\boldsymbol{\beta} \mid \mathbf{z}, \mathbf{y})\}\right], q^{(t)}_{MF}(\boldsymbol{\beta}) = \exp\left[E_{q^{(t-1)}_{MF}(z)}\{\log p(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{y})\}\right]$$

  which approximate $p(\boldsymbol{\beta}, \mathbf{z}, \mid y)$ via a multivariate Gaussian $q^*_{MF}(\boldsymbol{\beta})$ and a product of truncated normals $\prod_{i=1}^n q^*_{MF}(z_i)$.
- (Inconsistency results). The expectation of L_2 norm of posterior mean $\boldsymbol{\beta}$ with respect to the mean field posterior converges to zero; while it increases at the rate of $\sqrt{n}$ if the expectation is over the true posterior. Both assume $p \to \infty$.
- **(VB with partially factorized family)**. A better solution seems to find the solution within the family $q_{PMF}(\mathbf{z}, \boldsymbol{\beta}) = p(\boldsymbol{\beta} \mid \mathbf{z}) \prod_{i=1}^p q_{PMF}(z_i)$, and the solution guarantees of the convergence $KL[q^*_{PMF}(\boldsymbol{\beta})||p(\boldsymbol{\beta} \mid \mathbf{y})] \xrightarrow{p} 0$, when $p \to \infty$. **Computational scalability** for large p and large n?

Other approximation to multivariate normal CDF may include, e.g. low rank or sparse approximation of the covariance, and expectation propagation [Minka, 2013].

Extension

Answers to Q3:
There are a large number of applications and extensions, including:

- multivariate probit link,
- binary data with a latent nonlinear function modeled by a Gaussian process [Cao et al., 2022],
- model of binary time series by probit dynamic linear model [Fasano et al., 2021],
- skewed distribution as a more flexible class to use in approximation.

Other possible directions:

- Objective prior or objective choice of prior parameters.
- Computational scalable approaches for posterior credible interval of β.
- Variable selection when the number of covariates is large, or/and when coefficients are time varying.
- Approximation approaches on above extensions.

Thanks!

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. Journal of the American statistical Association, 88(422):669–679, 1993.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. Journal of the American statistical Association, 112(518):859–877, 2017.

Zdravko I Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(1):125–148, 2017.

Jian Cao, Daniele Durante, and Marc G Genton. Scalable computation of predictive probabilities in probit models with gaussian process priors. Journal of Computational and Graphical Statistics, pages 1–12, 2022.

Hao Chen, J Loeppky, Jerome Sacks, and W Welch. Analysis methods for computer experiments: How to assess and what counts? Technical Report 1, 2016.

Daniele Durante. Conjugate bayes for probit regression via unified skew-normal distributions. Biometrika, 106(4):765–779, 2019.

Augusto Fasano, Giovanni Rebaudo, Daniele Durante, and Sonia Petrone. A closed-form filter for binary time series. Statistics and Computing, 31 (4):1–20, 2021.

Thomas P Minka. Expectation propagation for approximate bayesian inference. arXiv preprint arXiv:1301.2294, 2013.